Data mining and analysis of bacillus virus

Jiwoo An[1], ChanWoo Kim[1], Seohyun Moon[1], Jungeun Huh[1] and Taeseon Yoon[2]

[1]*Natural Science, Hankuk Academy of foreign studies, Yong-in, Korea*
[2]*Computer Science, Korea University, Seoul, Republic of Korea*

Abstract. Anthrax- can be found naturally in soil and commonly affects domestic and wild animals around the world. Bacillus anthracis is mostly common viruses of Anthrax. There are some more occurred in similar DNA sequences: 3 Bacillus Viruses: Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis. This is a report about analyzing the similarities between the Bacillus viruses by investigating the frequency of amino acid and finding the difference between those three viruses based on the gene. Those diseases are infected by parasite and host animals, and cause muscle pain. Therefore, we can conclude that Leucine is a protein that plays a significant role in causing muscle pain. Secondly, In Analysis of decision tree, there are only little differences between each classes. The classes represent positions that include representative protein. The very position that windows mention is their difference.

1 Introduction

Bacillus, a bacteria, is a bar and cylindrical shaped aerobacter. There are a 34 species of bacillus virus, such as bacillus cereus, which causes food decay and has flagellum but doesn't have any capsule (different from bacillus anthracis), and bacillus thuringiensis, which produce sterilized toxin. However, most of them are avirulence, but bacillus anthracis is the only virus that causes anthrax. Therefore, Bacillus anthracis was utilized for a number of biochemical weapons. Since each bacillus virus causes different influence to the nature, we attempt to reveal the difference between bacillus cereus, bacillus thuringiensis and bacillus anthracis. We use decision tree program and apriori program as classification standards. Apriori is used to measure the frequency while the decision tree classifies elements that are unique.

2 Apriori experiment

We did an experiment about three bacillus viruses: bacillus anthracis, bacillus cereus, and bacillus thuringiensis. We made the results using apriori algorithm and decision tree algorithm. Before having an apriori experiment, we divided three bacillus into three classes: 13window, 17window, and 19window. Since the lengths of each bacillus are different, we divided them into different windows according to their lengths that the comparison with the viruses' position similar is possible when the sequences are repeated. We got similar consequences from each cases, thus overall we could conclude that acid I-Leucine has the highest frequency and that all three cases are involved in the actions occurring in the body.

2.1 Captions/numbering

We use apriori algorithm to figure out which protein has highest frequency in each of three bacillus viruses. Apriori algorithm is used to find out the most frequent protein, like frequent item set mining algorithm, and association rule. learning over transactional databases. So, we use this algorithm to know which amino acid mostly consists bacillus-Leucine-, why Leucine shows the highest frequency and how Leucine affects bacillus and our bodies.

2.2 Apriori results

We got exactly same results in all experiment-all windows of three viruses: Leucine.

Table 1. All experiment-all windows of three viruses

virus/ windows	Bacillus anthracis	Bacillus cereus	Bacillus thuringiensis
13 windows	I-Leucine (13850)	I-Leucine (13989)	I-Leucine (13925)
17 windows	I-Leucine (10591)	I-Leucine (10697)	I-Leucine (10649)
19 windows	I-Leucine (9476)	I-Leucine (9571)	I-Leucine (9528)

Numbers of Table 1. represent the average figure of Leucine found in each Amino groups.

Following Tables 2, 3, 4 are each about frequency analysis about Leucine in bacillis anthracis, bacillus cereus, and bacillus thuringiensis at 19 windows.

Table 2. Analysis of bacillus anthracis

position	window		
	13 (count)	*17 (count)*	*19(count)*
Position 1	L(13815)	L(10542)	L(9541)
position 2	L(13798)	L(10586)	L(9609)
position 3	L(13908)	L(10585)	L(9372)
position 4	L(13930)	L(10391)	L(9345)
position 5	L(13813)	L(10813)	L(9302)
position 6	L(13849)	L(10469)	L(9538)
position 7	L(13837)	L(10649)	L(9490)
position 8	L(13967)	L(10600)	L(9433)
position 9	L(13943)	L(10635)	L(9590)
position 10	L(13897)	L(10509)	L(9514)
position 11	L(13679)	L(10633)	L(9515)
position 12	L(13763)	L(10714)	L(9488)
position 13	L(13849)	L(10533)	L(9524)
position 14		L(10509)	L(9486)
position 15		L(10683)	L(9482)
position 16		L(10516)	L(9382)
position 17		L(10672)	L(9390)
position 18			L(9636)
position 19			L(9402)

Table 3. Analysis of bacillus cereus

position	window		
	13 (count)	*17 (count)*	*19(count)*
position1	L(14030)	L(10713)	L(9519)
position 2	L(13937)	L(10828)	L(9576)
position 3	L(14040)	L(10537)	L(9715)
position 4	L(13983)	L(10700)	L(9538)
position 5	L(13906)	L(10688)	L(9541)
position 6	L(13975)	L(10625)	L(9586)
position 7	L(13955)	L(10630)	L(9696)
position 8	L(14077)	L(10782)	L(9592)
position 9	L(13943)	L(10679)	L(9523)
position 10	L(14207)	L(10612)	L(9546)
position 11	L(13885)	L(10753)	L(9569)
position 12	L(14004)	L(10889)	L(9561)
position 13	L(13914)	L(10841)	L(9574)
position 14		L(10623)	L(9719)
position 15		L(10668)	L(9387)
position 16		L(10535)	L(9685)
position 17		L(10753)	L(9548)
position 18			L(9471)
position 19			L(9510)

2.2.1 Results analysis

As you see in the tables ahead, results of all three bacillus is same-Leucine. Even though there are small differences among figures-exactly show how much time Leucine had been detected-, such things are so insignificant that it cannot cause substantive distinction inside the organism. Therefore, we could figure out that Leucine plays important role in most bacillus virus. For instances, most effects of amino acids on bacillus signaling are abolished by lowering the concentration of Leucine.

Table 4. Analysis of bacillus thuringiensis

position	window		
	13 (count)	*17 (count)*	*19(count)*
position1	L(13967)	L(10492)	L(9652)
position 2	L(14027)	L(10554)	L(9365)
position 3	L(13942)	L(10753)	L(9570)
position 4	L(13919)	L(10592)	L(9589)
position 5	L(13869)	L(10743)	L(9476)
position 6	L(14106)	L(10772)	L(9361)
position 7	L(13750)	L(10679)	L(9480)
position 8	L(13904)	L(10676)	L(9452)
position 9	L(13877)	L(10667)	L(9455)
position 10	L(13976)	L(10616)	L(9595)
position 11	L(13870)	L(10597)	L(9647)
position 12	L(13925)	L(10696)	L(9500)
position 13	L(13898)	L(10649)	L(9624)
position 14		L(10613)	L(9652)
position 15		L(10657)	L(9468)
position 16		L(10708)	L(9474)
position 17		L(10566)	L(9505)
position 18			L(9607)
position 19			L(9558)

2.2.2 Leucine and myalgia (How Leucine effects on our bodies)

Bacillus anthracis, which is infectious to animals, is endemic(Anthrax)-a disease which regularly transmit infection to regional societies-. In some cases, it is infected by physical contacts, thus mostly it occurs in certain region or tribes. You would suffer from myalgia, serious cough, dyspnea, papule, and so on. Not only anthrax but also generous endemics-such as malaria, dengue fever, yellow fever, typhoid fever, cholera, etc-contain those symptoms, especially myalgia. And apriori experiment results of those viruses are the same: Leucine. Therefore, we may concluded that Leucine consists most of endemic viruses and brings out myalgia.

The experiment is adapting Quinla's C 5.0 [3] algorithm and rule extraction method. See 5.0 program is used for the experiment.

Fig. 1 shows abstract shape of common dicision tree, which we used, and Fig. 2 shows the part of decision tree and condition to get to the next node at particular situation.

3 Decision tree

3.1 Decision tree algorithm

We use Decision tree algorithm to find out the difference between three bacillus viruses. Decision tree algorithm showed position of amino acid. So, we used this algorithm to know how can we classify best according to position. Thus, we utilized this algorithm in order to figure out how we can classify the best regarding to position.

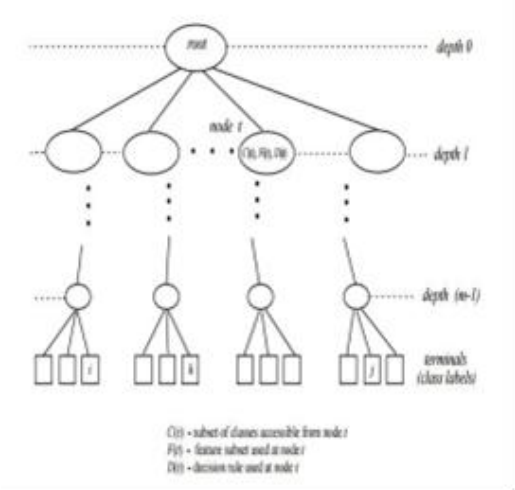

Figure 1. manual form of decision tree [1]

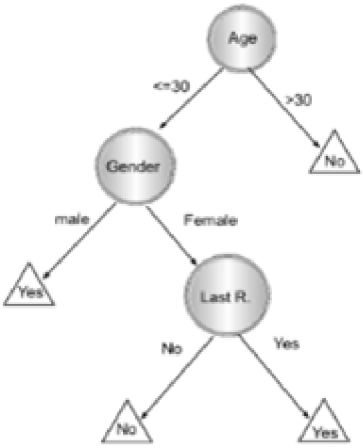

Figure 2. Decision tree presenting response to direct mailing.[2]

3.2 Decision tree

There are mainly two types of decision trees according to usage. Classification tree which is used for classification tasks and regression tree, used for regression tasks. [4]

Our usage of the decision tree is the former one, to classify a case by starting from the tree root and moving it till a leaf is encountered. The case's result for the test at individual non-leaf decision node is decided and focus moves to the subtree's root, matching to this outcome. After this process ultimately (and naturally) generates to a leaf, it is expected that the class of the case to be that noted at the leaf. [3]

A path is made between the root of the decision tree and one of its leaves when the tree is used to classify a case. To reach a single leaf, it must accord with entire conditions on the path. We can know that because of this, decision tree is highly related to rule induction [5]

3.3 Decision tree results

According to the Table 1, we can see that those 3 viruses have a big similarity. Which means, it is hard to find out specificity at each virus. By this result, we can assume that all of these viruses are derived from one same virus and there are small proportion of mutein. According to Table 5, we have to notice rule extraction under 13 window showed their rules with amino acid at position 11. Glutamic Acid (E), Glysin(G), are Isoleucine(I) are all presented on 11^{th} position in 13 window, and this common position represents the property of bacillus virus. Moreover, second position is also playing an important role, presenting G on class 1, N on class 2, and G and C on class 3. In Table 6, we have to notice that under 17 window, rules were extracted at position 16. At Table 7, we can find out under 19 window, position 8 was the place where rules were extracted. Also, we can see that those position played important role differentiating each subtype from each other.

Following Tables 5, 6, 7 are rule extraction due to the position of amino acid.

Table 5. 13 windows

Virus	**Rule**		
Class 1	position2=G	position10=I	position11=E
	position2=L	position11=G	
	position3=V	position11=I	
	position4=I	position11=F	
	position6=W	position11=L	
Class 2	position2=N	position10=P	position11=E
	position2=A	position11=G	
	position3=H	position11=I	
	position4=A	position11=F	
Class3	position2=G	position10=V	position11=E
	position2=C	position11=G	
	position3=S	position11=I	
	position4=M	position11=F	
	position6=E	position11=L	

Table 6. 17 windows

Virus	**Rule**	
Class 1	position8=R	position16=I
	position7=S	position8=C
	position4=G	position8=A
	position8=W	position14=E
Class 2	position8=R	position16=S
	position7=K	position8=C
	position4=E	position8=A
	position8=W	position14=L
Class 3	position8=R	position16=Q

	position7=H	position8=C
	position8=W	position14=F

Table 7. 19 windows

Virus	**Rule**	
Class 1	position4=H	position16=A
	position5=Q	position16=E
	position1=M	position16=V
	position10=Y	position16=N
Class 2	position4=Q	position16=A
	position5=S	position5=E
	position1=G	position16=V
	position10=Q	position16=N
Class 3	position1=H	position16=V
	position4=C	position16=A

4 Conclusion

Firstly, In Analysis of apriori, Three kinds of bacillus virus have Leucine the most (among.....). Leucine is a virus detected in endemic diseases such as malaria [6], dengue fever [7], typhoid [8], serious cough, dyspnea, papule and so on [9,10]. Those diseases are infected by parasite and host animals, and cause muscle pain. Therefore, we can conclude that Leucine is a protein that plays a significant role in causing muscle pain. Secondly, In Analysis of decision tree, there are only little differences between each classes. The classes represent positions that include representative protein. The very position that windows mention is their difference. We are looking forward to conceive natural treatment(not the artificial one) to cure endemic disease. Using common property of bacillus virus that it has Leucine, which is contained in endemic disease, we are planning to analyze the similarity and develop the natural treatment.

References

1. Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." *IEEE transactions on systems, man, and cybernetics* **21.3** (1991): 660-674.
2. Rokach, Lior, and Oded Maimon. "Top-down induction of decision trees classifiers-a survey." *Systems, Man, and Cybernetics*, Part C: Applications and Reviews, IEEE Transactions on **35.4** (2005): 476-487.
3. Quinlan, J. Ross. C4. 5: programs for machine learning. Elsevier, 2014.
4. L. Rokach and O. Maimon, Data Mining with Decision Trees: *Theory and Applications, World Scientific Pub Co Inc.*, (2008).
5. Quinlan, J. Ross. "Simplifying decision trees." *International journal of man-machine studies* **27.3** (1987): 221-234.
6. Lee, Jung-Yub, Su-Min Song, Ji-Woong Seok, Bijay Kumar Jha, Eun-Taek Han, Hyun-Ouk Song, Hak-Sun Yu, Yeonchul Hong, Hyun-Hee Kong, and Dong-Il Chung. "M17 Leucine Aminopeptidase of the Human Malaria Parasite Plasmodium Vivax." *Molecular and Biochemical Parasitology* **170.1** (2010): 45-48.
7. Fatima, Zareen, Muhammad Idrees, Mohammad A. Bajwa, Zarfishan Tahir, Obaid Ullah, Muhammad Q. Zia, Abrar Hussain, Madiha Akram, Bushra Khubaib, Samia Afzal, Saira Munir, Sana Saleem, Bisma Rauff, Sadaf Badar, Mahrukh Naudhani, Sadia Butt, Mahwish Aftab, Liaqat Ali, and Muhammad Ali. "Serotype and Genotype Analysis of Dengue Virus by Sequencing Followed by Phylogenetic Analysis Using Samples from Three Mini Outbreaks-2007-2009 in Pakistan." *BMC Microbiology BMC Microbiol* **11.1** (2011): 200.
8. Kariuki, Samuel, Gunturu Revathi, John Kiiru, Doris M. Mengo, Joyce Mwituria, Jane Muyodi, Agnes Munyalo, Yik Y. Teo, Kathryn E. Holt, Robert A. Kingsley, and Gordon Dougan. "Typhoid in Kenya Is Associated with a Dominant Multidrug-Resistant Salmonella enterica Serovar Typhi Haplotype That Is Also Widespread in Southeast Asia." *Journal of Clinical Microbiology. American Society for Microbiology (ASM)*, n.d. Web. 29 Sept. (2015).
9. "Cholera: Causes, Symptoms, Treatment, and Prevention." *WebMD. WebMD*, n.d. Web. 29 Sept. (2015).
10. "Anthrax Causes, Symptoms, Treatment - Anthrax Signs and Symptoms - EMedicineHealth." *EMedicineHealth. N.p.*, n.d. Web. 29 Sept. (2015).